Testing the biases in the rich Cenozoic Angiosperm macrofossil record

Xing, Yaowu ; Gandolfo, Maria A ; Onstein, Renske E ; Cantrill, David J ; Jacobs, Bonnie F ; Jordan, Gregory J ; Lee, Daphne E ; Popova, Svetlana ; Srivastava, Rashmi ; Su, Tao ; Vikulin, Sergei V ; Yabe, Atsushi ; Linder, H Peter

Abstract: Premise of research: The Cenozoic fossil record is crucial for understanding the evolution of the remarkably high diversity of angiosperms. However, the quality and biases of the angiosperm fossil record remain unclear mainly due to the lack of a global database. Methodology: We introduce a new global occurrence-based database for Cenozoic angiosperm macrofossils, the Cenozoic Angiosperm Database. We test the temporal, spatial, and phylogenetic biases of the Cenozoic angiosperm macrofossil record and explore their causes. Pivotal results: The data presented here include 2478 assemblages from all Cenozoic epochs and 1961 sites from all continents, as well as representatives of 221 families (of 445 recognized) and 1859 genera, and show that the Cenozoic angiosperm macrofossil record is extraordinarily rich. However, this rich record is temporally, spatially, and phylogenetically biased: the Miocene is much better sampled than the rest of Cenozoic, the Northern Hemisphere is better sampled than the Southern Hemisphere, and the rosids are better sampled than the rest of the angiosperms. The sampling bias might be caused by collecting effort, geological history, or diverse features of the families, such as growth form and distribution. Conclusions: The Cenozoic macrofossil record of angiosperms is remarkably rich, especially of woody families found in the Northern Hemisphere. Even if there are numerous biases in these data, a judicious use of the database should be highly informative.

DOI: https://doi.org/10.1086/685388

Int. J. Plant Sci. 177(4):000–000. 2016.

1058-5893/2016/17704-00XX$15.00 DOI: 10.1086/685388

TESTING THE BIASES IN THE RICH CENOZOIC ANGIOSPERM MACROFOSSIL RECORD

Yaowu Xing,[1,*] Maria A. Gandolfo,† Renske E. Onstein,* David J. Cantrill,‡ Bonnie F. Jacobs,§ Gregory J. Jordan,‖ Daphne E. Lee,# Svetlana Popova,** Rashmi Srivastava,†† Tao Su,‡‡ Sergei V. Vikulin,** Atsushi Yabe,§§ and H. Peter Linder*

*Institute of Systematic Botany, University of Zürich, Zollikerstrasse 107, 8008 Zürich, Switzerland; †L. H. Bailey Hortorium, Plant Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, New York 14853, USA; ‡Royal Botanic Gardens Melbourne, Private Bag 2000, Birdwood Avenue, South Yarra, Victoria 3141, Australia; §Environmental Science Program, Southern Methodist University, PO Box 750395, Dallas, Texas 75275-0395, USA; ‖School of Plant Sciences, University of Tasmania, Private Bag 55, Hobart, Tasmania 7001, Australia; #Department of Geology, University of Otago, Dunedin 9016, New Zealand; **Komarov Botanical Institute, Russian Academy of Sciences, 197376 Saint Petersburg, Russian Federation; ††Birbal Sahni Institute of Palaeobotany, Lucknow 226 007, India; and ‡‡Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla 666303, China; and §§National Museum of Nature and Science, 4-1-1 Amakubo, Tsukuba 305-0005, Japan

Editor: Maria von Balthazar

Premise of research. The Cenozoic fossil record is crucial for understanding the evolution of the remarkably high diversity of angiosperms. However, the quality and biases of the angiosperm fossil record remain unclear mainly due to the lack of a global database.

Methodology. We introduce a new global occurrence-based database for Cenozoic angiosperm macrofossils, the Cenozoic Angiosperm Database. We test the temporal, spatial, and phylogenetic biases of the Cenozoic angiosperm macrofossil record and explore their causes.

Pivotal results. The data presented here include 2478 assemblages from all Cenozoic epochs and 1961 sites from all continents, as well as representatives of 221 families (of 445 recognized) and 1859 genera, and show that the Cenozoic angiosperm macrofossil record is extraordinarily rich. However, this rich record is temporally, spatially, and phylogenetically biased: the Miocene is much better sampled than the rest of Cenozoic, the Northern Hemisphere is better sampled than the Southern Hemisphere, and the rosids are better sampled than the rest of the angiosperms. The sampling bias might be caused by collecting effort, geological history, or diverse features of the families, such as growth form and distribution.

Conclusions. The Cenozoic macrofossil record of angiosperms is remarkably rich, especially of woody families found in the Northern Hemisphere. Even if there are numerous biases in these data, a judicious use of the database should be highly informative.

Keywords: angiosperm fossil record, fossil database, taphonomic bias, temporal bias, fossil quality, fossil completeness.

Online enhancement: appendix.

Introduction

Fossils, as direct evidence of past life, have long been used for inferring diversity dynamics (Raup 1972; Sepkoski et al. 1981; Raup and Sepkoski 1982; Fröbisch 2013; Pearson et al. 2013). However, accurate estimation of past diversity relies on having a reliable fossil record. Paleontologists have long recognized that the fossil record is incomplete and biased by taphonomy, sampling, eustasy, and tectonic processes (Adrain and Westrop 2003). Therefore, measuring the biases of the existing fossil record and being able to take these biases into account will improve our understanding of past diversity over long timescales (Durham 1967; Benton and Storrs 1994; Benton et al. 2000; Foote et al. 2015).

Many methods have been proposed to test the quality or completeness of the fossil record. These methods are based on gaps in the stratigraphic ranges of fossil taxa, on hypothetical lineages implied by estimated evolutionary trees, or on estimates of the probability of genus preservation per stratigraphic interval (Maxwell and Benton 1990; Benton and Storrs 1994; Kidwell and Flessa 1995; Foote and Sepkoski 1999; Benton et al.

[1] Author for correspondence; current address: Field Museum of Natural History, 1400 South Lake Shore Drive, Chicago, Illinois 60605, USA; e-mail: yxing@fieldmuseum.org.

Manuscript received July 2015; revised manuscript received November 2015; electronically published April 6, 2016.

2000). Most previous studies have focused on various groups of vertebrates (Benton et al. 2011) and insects (Karr and Clapham 2015). Bias testing on plant fossil records is scarce. Few studies have tested the fossil quality of Phanerozoic vascular plants (Cascales-Miñana et al. 2010; Cascales-Miñana and Diez 2012), the Paleozoic plant fossil record (Cascales-Miñana 2011), and gymnosperms (Cascales-Miñana and Cleal 2013), but specific bias tests have not been applied to angiosperms.

The angiosperms, with a standing diversity of approximately 350,000 species (Paton et al. 2008), represent by far the largest clade of terrestrial plants. How angiosperms achieved such high diversity has long challenged biologists, but clues may be found in their Cenozoic diversification history. Fossil-based analyses of the patterns of angiosperm evolution for the Cretaceous have led to important insights into the initial disparification and species radiation of the group (Lidgard and Crane 1988, 1990; Crane and Lidgard 1989; Lupia et al. 1999). However, the diversification of angiosperms during the Cenozoic, and the causes of such changes in diversity, remains unclear. Undoubtedly, the spatial and temporal distribution of angiosperm fossils will contribute substantially to our understanding of angiosperm evolutionary history, life habits, and the environments in which they lived and reproduced. Even though thousands of fossil angiosperm species have been described and published, very little is known about the quality, completeness of, and biases in the existing Cenozoic angiosperm fossil record.

Here we present a new global occurrence-based database for the Cenozoic angiosperm macrofossils, the Cenozoic Angiosperm Database (CAD), and discuss its phylogenetic, taxonomic, spatial (by continent), and temporal (by epoch) coverage. We investigate possible spatial biases in the CAD caused by the accessibility of fossil localities, the intensity of study (Northern Hemisphere vs. Southern Hemisphere), and the continental erosional history, as well as phylogenetic biases caused by different preservation rates. We also evaluate the completeness of the Cenozoic angiosperm macrofossil record with the hope that this database will ultimately provide a fresh view of Cenozoic diversity patterns of the angiosperms.

Material and Methods

Data Compilation and Accessibility

The CAD is an occurrence-based database, making it possible to infer absences of taxa from time intervals falling within their age ranges (Lazarus taxa). Second, it makes it possible to calculate the frequency with which taxa are encountered in each time interval (Alroy 2010). Finally, it is easy to curate, update, and add new occurrences.

The CAD is structured into five relational tables. The Site Table contains geographical information for each fossil assemblage. The Assemblage Geology Table encompasses the information on the name, age, epoch, and stages of the formation and also additional information on the assemblages and the fossil forms (such as leaves, stems, etc.). The Assemblage Taxon Table comprises information specific to each taxon, including identification reliability, nearest living relatives, and growth forms where available. The taxonomy of each genus is presented in the Taxonomy Table, where fossil genera are assigned to higher taxonomic levels such as family, order, and so on, following the APG III system (Angiosperm Phylogeny Group 2009). All references are compiled in the Reference Table.

The intention is to include all published angiosperm macrofossils from the Cenozoic. Under macrofossils we include wood, leaf, seed, fruit, flower, inflorescence, and dispersed cuticle fossils. We prioritized macrofossils over microfossils because of better determination of taxonomic position. However, the structure of the CAD will make it easy to incorporate the paleopalynological data. Few Pleistocene records are represented in the CAD, mainly because most Pleistocene angiosperm fossils have been overlooked by paleobotanists, as they are comparable with extant taxa.

The data in the CAD were compiled in two ways: first, by combining data from two existing databases, the Paleobiology Database (http://paleodb.org, accessed June 20, 2011) and the Palaeoflora Database (Utescher and Mosbrugger 2012), and, second, by manually entering data from the literature published between 1836 and 2014 (fig. 1; appendix, available online). The data from other databases were modified to fit the CAD structure, and, where possible and necessary, the related information was upgraded (e.g., all sites are georeferenced and the taxonomy was upgraded to APG III; Angiosperm Phylogeny Group 2009).

The CAD is freely accessible at http://www.fossil-cad.net, which was registered on March 10, 2015. A user-friendly interface facilitates searches of the database and assists in the download of data. Anyone who wants to contribute data can contact Y. Xing by e-mail.

Fossil Assemblage and Site Data

Assemblages are the basic units in the CAD. In most cases a macrofossil assemblage can be considered the equivalent of

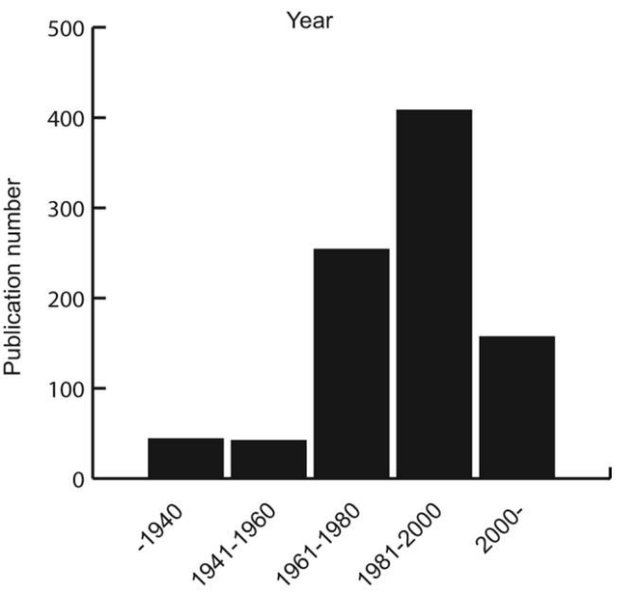


Fig. 1 Publications included in the Cenozoic Angiosperm Database.

a modern community; it is a group of species (fossil occurrences) recorded from a particular time (bed) and place. Within the CAD, the concept of assemblage is the equivalent of a taphoflora sensu Romero (1986), which is the list of the fossil taxa preserved in a given outcrop. The taxa are representatives of the original flora as modified by taphonomic processes (such as dispersal, burial site, and diagenesis) and taxonomic biases (the presence of diagnostic characteristics on the fossils, level of knowledge of relevant floras, and time constraints on the paleobotanists).

"Site" refers to the geographical location of the assemblage. Each site has unique coordinates that were taken, when available, from the original publication. However, if precise coordinates were not included in the original publication, the coordinates were obtained from Google Earth. In the context of this database, a geological formation can contain several sites, and a site can contain several assemblages with different geological ages. Cenozoic fossil sites are assigned to seven continents or regions: Africa, Antarctica, Asia, Australia/Oceania, Europe, North America, and South America.

We followed the Geologic Time Scale 2012 (Gradstein et al. 2012), and ages within different geological frameworks were translated to this system. Ages of some assemblages were updated when more accurate determination was available. For example, Gelasian assemblages, now included in the Pleistocene, were considered of Pliocene age in most literature. They were updated into Pleistocene if exact age was available. However, in those particular cases that the Gelasian data cannot be recognized, they were assigned to the Pliocene.

Taxonomic Data

Cenozoic angiosperm data are included in the CAD irrespective of the quality of the identifications. This was done to maximize the number of records of rare species and to avoid arbitrary decisions as to identification accuracy. For most fossil occurrences, only the original identifications are available, and relatively few fossils have been reevaluated by experts. Regrettably, this means that there are many incorrectly determined fossils in the database. Users should consider the identification error in the CAD and apply available methods to reduce this bias. Our use of the APG III (Angiosperm Phylogeny Group 2009) classification to place the fossil taxa within families, orders, and larger groupings means that a putative phylogenetic placement of the fossils is reflected in the CAD.

Tests of Sampling Biases

Sampling bias is a major problem when addressing patterns of taxonomic diversity at all scales, and the plant fossil record may be affected by biases in taxon sampling, preservation, and identification. Consequently, we tested for temporal, continental, phylogenetic, and growth-form sampling biases.

Temporal biases are here understood to be the consequence of the variable duration of the geological epochs. It is evident that more fossil sites are likely to be found from longer time intervals, and these consequently inflate the estimated diversity: we refer to this as the "timescale bias." We used linear regression to test whether the number of fossil sites or assemblages is predicted by the duration of the epochs. We also set up an alternative timescale with part-epochs of more or less equal duration to test whether this removed the timescale bias.

Spatial biases may have two causes. First, continents differ in land area and exposed sedimentary rock. Globally, sedimentary rocks cover approximately 65%–67% of all land surface (Blatt and Jones 1975; Ronov 1982; Amiotte Suchet et al. 2003). Overall, there is less exposed sedimentary rock in the continents of the Southern Hemisphere (Africa, Antarctica, Oceania, South America) than in the Northern Hemisphere (Asia, Europe, North America; table 1). We used regression analysis to test whether the number of fossil sites and assemblages is predicted by the land and sedimentary area of each continent. As the proportion of exposed sedimentary rocks is unavailable for Antarctica, it is excluded in the latter analysis. Second, paleofloras in the Northern Hemisphere continents have been more intensively studied than paleofloras in the Southern Hemisphere. In order to test whether families currently restricted to the Northern Hemisphere are more likely to be present in the fossil record than families restricted to the Southern Hemisphere, we set a null hypothesis that Northern and Southern Hemisphere families are equally likely to be found in the record. We compared the observed presence in the fossil record against this null with contingency tests (χ^2 and Fisher exact tests). The continental distribution of the families was based on the Angiosperm Phylogeny website (http://www.mobot.org/, accessed August 1, 2014), and cosmopolitan families were ignored in this analysis.

Phylogenetic biases may result from different probabilities of preservation through the interaction of anatomy and taphonomy (Kidwell and Holland 2002). We used two methods to test for phylogenetic bias in the fossil record. First, we asked whether fossilization shows a phylogenetic signal and tested whether the families with fossil record, as compared to families without fossil record, were randomly distributed on a time-calibrated angiosperm family-level phylogeny (Qian and Zhang 2014). We generated a distribution test by randomizing the presence/absence data 1000 times over the phylogeny and calculating for each the number of evolutionary steps needed to place the data on the tree. We compared this to the observed number of steps. If fossilization is phylogenetically random, then the observed number of steps should lie within 95% of the step number generated from the randomized data. This approach does not take into account variation in branch length or any probability of fossilization (larger family size, many woody species, etc.). In the second analysis, we assigned the number of occurrences in the fossil record for each family to the family-level tips of the tree and used Blomberg's *K*, implemented in the "phylosignal" function in the R package "phytools" (Revell 2013), to test whether observed variance differs significantly from that derived from the given phylogeny under Brownian motion (Blomberg et al. 2003). There is phylogenetic signal if Blomberg's *K* is significantly greater than 0.

To disentangle the causes of the taphonomic biases, we tested whether there is support for correlation between the different growth forms (mainly herbaceous vs. mainly woody), distribution areas (Northern Hemisphere, Southern Hemisphere, or both) and distribution ranges (in one to six continents), and presence/absence in the fossil record on the phylogeny. For each potential explanatory variable, we compared an independent (uncorrelated) and dependent (correlated) model using the function DISCRETE in BayesTraits v 2.0 (Pagel and Meade

Table 1

Number of Fossil Sites and Assemblages per Continent and Epoch

	Area	Sedimentary rocks	Paleocene	Eocene	Oligocene	Miocene	Pliocene	Total
Africa	30.37	17.61	4, 4	5, 5	6, 6	29, 29	19, 34	63, 78
Antarctica	13.72	NA	2, 2	4, 4	2, 2	1, 1	1, 1	9, 9
Asia	43.82	32.43	67, 70	97, 108	103, 225	256, 417	91, 100	614, 920
Europe	10.18	8.86	8, 8	35, 36	83, 94	284, 377	31, 31	441, 546
North America	24.49	12.73	426, 448	187, 241	21, 21	45, 46	15, 15	694, 771
Oceania	9.01	6.31	4, 4	20, 30	15, 15	23, 24	6, 6	68, 79
South America	17.84	11.06	3, 3	6, 6	6, 6	49, 52	8, 8	72, 75
Total	149.43	89	514, 539	354, 430	236, 369	687, 946	171, 195	1961, 2478

Note. Data for epochs are shown as sites, assemblages. Areas of continents and exposed sedimentary rocks (million km^2) are also given.

2006). We used a reversible jump–Markov chain Monte Carlo model with priors obtained from the hyperprior approach and used an exponential prior seeded from a uniform distribution from the interval value of 0–30. All analyses were run for 5 million generations and replicated five times. The burn-in was 10%, and models and rate parameters were sampled every one-hundredth iteration. Difference between the models was assessed by Bayes factors (BFs), in which log BF = 2 × (log[harmonic mean(dependent model)] − log[harmonic mean(independent model)], where the harmonic mean is approximated by the harmonic mean of likelihoods averaged from the five runs. The model with the highest harmonic mean of likelihoods is regarded as the better one, and a logarithm BF value of 2–5 is interpreted as positive, 5–10 as strong, and >10 as highly significant (Pagel and Meade 2006). Support for the dependent model would suggest that the evolutionary transitions from absence to presence (or vice versa) in the fossil record are dependent on the state the other character is in, e.g., woody or herbaceous.

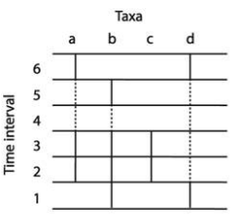
Completeness Tests

The simple completeness metric (SCM) test (Benton 1987) was used to quantify the relative completeness of the Cenozoic angiosperm fossil record. This metric is the ratio of observed fossil occurrences to total inferred fossil occurrences, as illustrated in figure 2. Total inferred fossil occurrences are calculated as observed fossils plus Lazarus taxa (Fara and Benton 2000). Lazarus taxa are postulated from a time interval without a record of that taxon separated by two or more time intervals. Theoretically, that taxon should also occur in the missing time interval.

In order to reduce the sampling bias per epoch, the Cenozoic was separated into eight time intervals with relatively equal durations (i.e., Paleocene: 65.5–55.8 Ma, 528 assemblages; early Eocene: 55.8–48.6 Ma, 204 assemblages; middle Eocene: 48.6–37.2 Ma, 54 assemblages; late Eocene: 37.2–33.9 Ma, 51 assemblages; Oligocene: 33.9–23.0 Ma, 352 assemblages; early Miocene: 23.0–16.0 Ma, 189 assemblages; middle Miocene: 16.0–11.6 Ma, 308 assemblages; late Miocene–Pleistocene: 11.6–0.1 Ma, 343 assemblages). This analysis included 215 families when assemblages with uncertain ages and taxa with uncertain phylogenetic positions were excluded. The SCM for each family was calculated for each time interval. As the Late Cretaceous data were not included, the SCM of Paleocene could not be assessed. We calculate the SCM of the late Miocene–Pleistocene by comparison between the middle Miocene and the Present. We used linear regression to test whether the SCM of each time interval is predicted by the number of assemblages.

In order to test whether there was a phylogenetic bias in the completeness for the major angiosperm groups (i.e., the basal angiosperms, the magnoliids, the monocots excluding Commelinids, the Commelinids, the rosids, the asterids, and the basal eudicots), we calculated the mean SCM of all families included in each group. The SCM for families not in the fossil record was scored as 0.

Plants with different growth forms and in different habitat have different preservation probabilities (Behrensmeyer and Kidwell 1985; Burnham and Spicer 1986; Greenwood 1991; Spicer 1991). In order to test whether the completeness is affected by plant growth form and habitat, we coded families into three categories: (*a*) mostly woody (hereafter "woody"), (*b*) aquatic (wetland) and mostly herbaceous (hereafter "wetland"), and (*c*) mostly nonaquatic herbaceous groups (hereafter "herbaceous"). The scoring is based on the abundance of fossil occurrence numbers in the CAD and not on the extant species numbers. For instance, the extant Primulaceae (s.l.) comprise woody and herbaceous species, and most of the herbaceous species are not aquatic. However, the most abundant genus of Primulaceae in the CAD, *Lysimachia*, containing more than 80% Primulaceae



	a	b	c	d	SCM
6	+	-	-	+	100%
5	-	+	-	-	33%
4	-	-	-	-	0%
3	+	+	+	-	75%
2	+	+	+	-	75%
1	-	+	-	+	100%
SCM	60%	80%	100%	30%	

Fig. 2 The simple completeness metric (SCM) and Lazarus taxa. *Left*, the stratigraphic ranges of four taxa (a–d) within six stratigraphic units (1–6). Solid lines indicate the presence of the taxa, and dotted lines represent their absence, illustrating the Lazarus effect. *Right*, calculation of the SCM, the ratio between the number of Lazarus units and the total number of units. The units can be either stratigraphic intervals (rows) or taxonomic groups (columns; after Fara and Benton 2000).

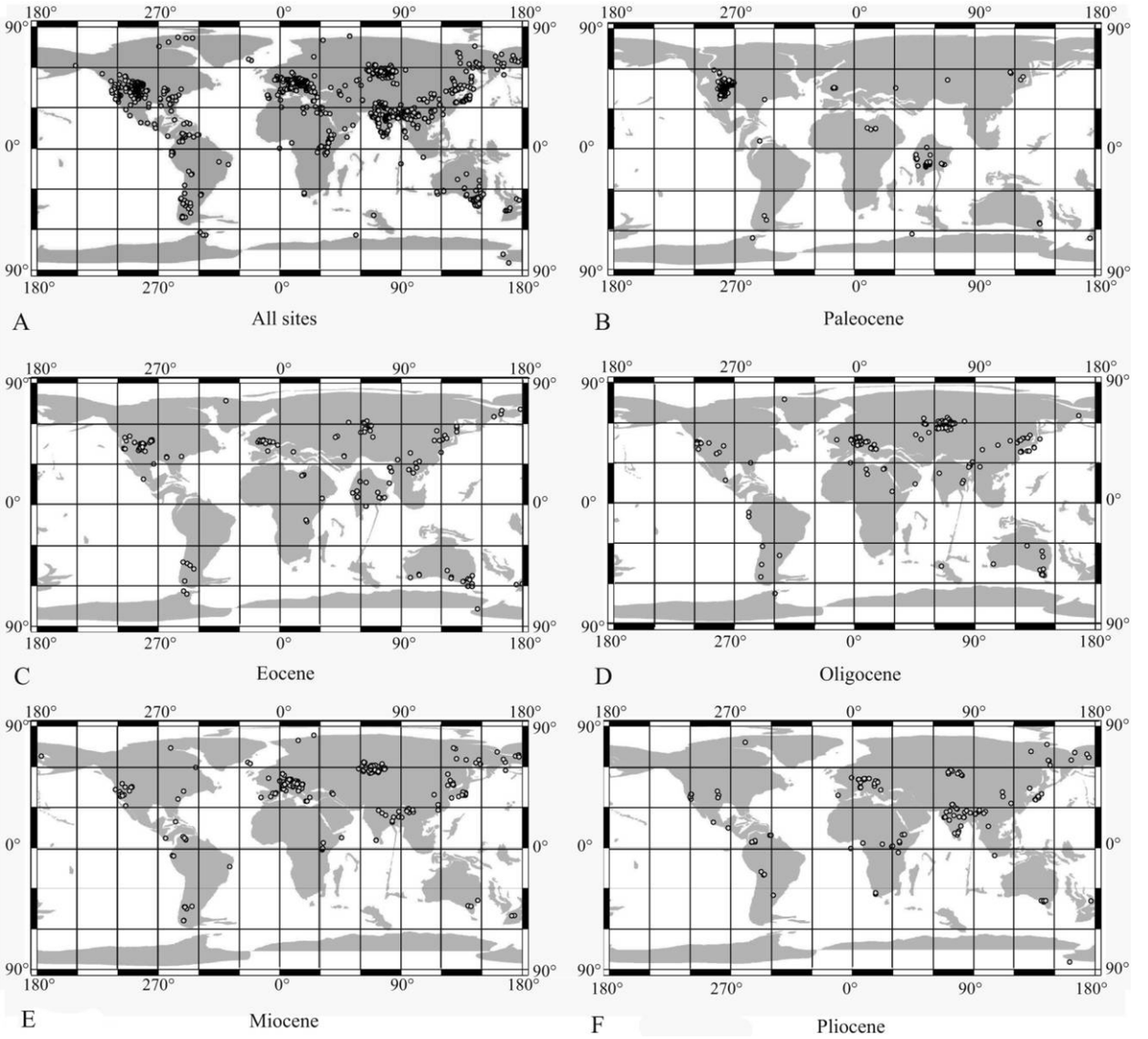


Fig. 3 Cenozoic angiosperm fossil sites in the Cenozoic Angiosperm Database. The paleocoordinates and figures are generated through the Ocean Drilling Stratigraphic Network Plate Tectonic Reconstruction Service (http://www.odsn.de/odsn/services/paleomap/paleomap.html). *A*, The whole Cenozoic; *B*, Paleocene; *C*, Eocene; *D*, Oligocene; *E*, Miocene; *F*, Pliocene.

occurrences, mainly occurs in wetlands, and, therefore, Primulaceae was scored as wetland. We calculated the occurrence number to check whether some growth forms are undersampled. The mean completeness for woody, wetland, and herbaceous groups at family level was calculated.

Results

Statistics of the CAD

As of March 5, 2015, the CAD contained data from 2478 assemblages representing 1961 sites (fig. 4; table 1). In total, this includes 49,965 Cenozoic angiosperm occurrences, of which 40,799 have been assigned at least to family level. The remaining taxa are incertae sedis, or taxa such as *Carpolithes*. For the Cenozoic, 220 angiosperm families, 1859 genera, and 9747 species are recorded, of which all families and 1172 genera (63%) are extant; 4495 species, belonging to 909 genera and 45 families, appear only once in the CAD.

Table 2

Linear Regressions of Number of Fossil Sites and Assemblages against Duration of Epochs, Area of Continents, and Area of Exposed Sedimentary Rocks

	Site no.		Assemblage no.	
	R^2	P	R^2	P
Continental area	.22	.28	.32	.18
Sedimentary rocks	.19	.39	.34	.23

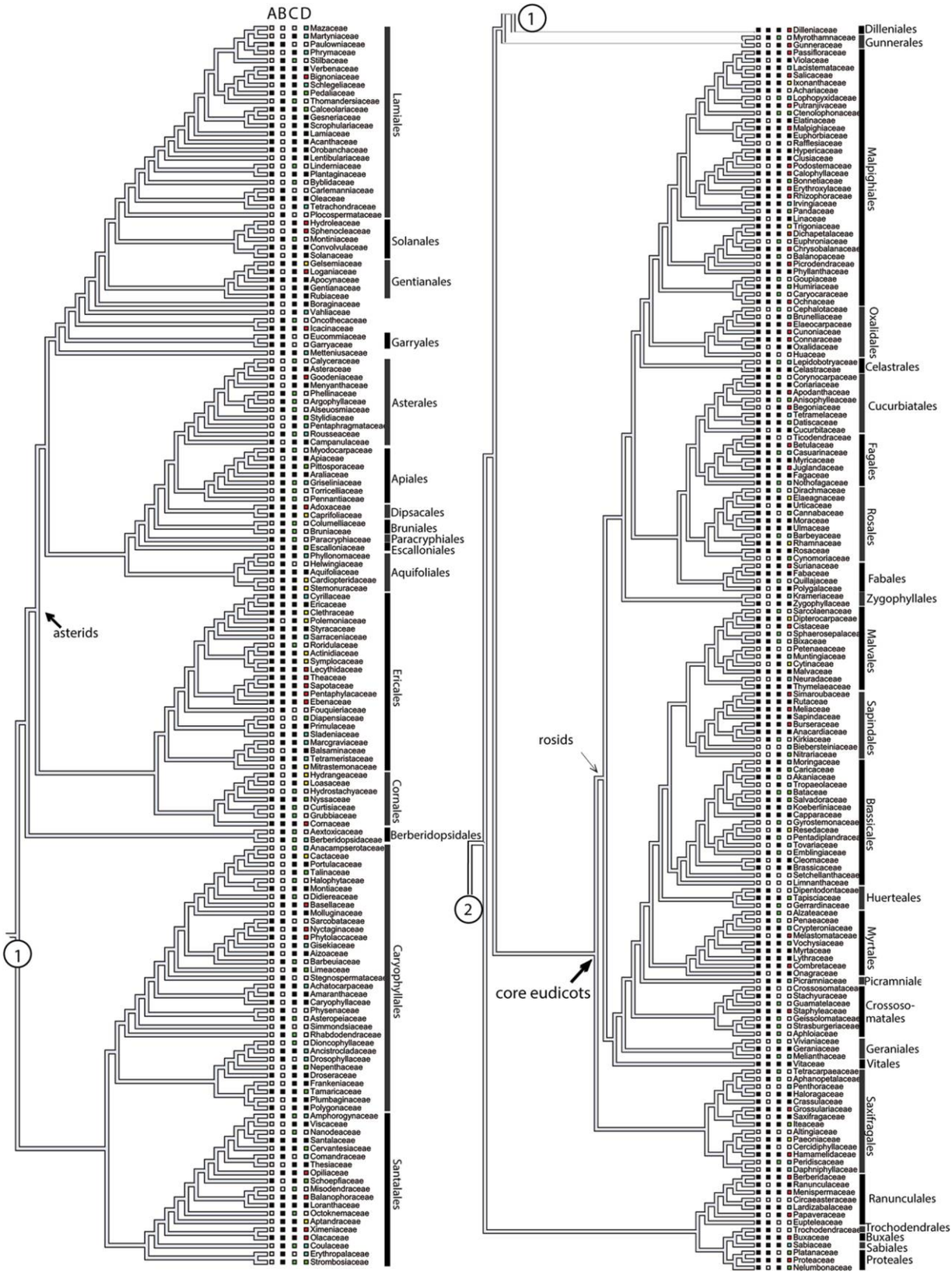


Fig. 4 Phylogenetic distribution of the Cenozoic angiosperm fossil record at family level and their habits, distribution area (hemisphere), and ranges (number of continents). The family-level phylogenetic tree is based on phylogeny of Qian and Zhang (2014). Numbers in circles indicate where subsets of the phylogenetic tree are connected.

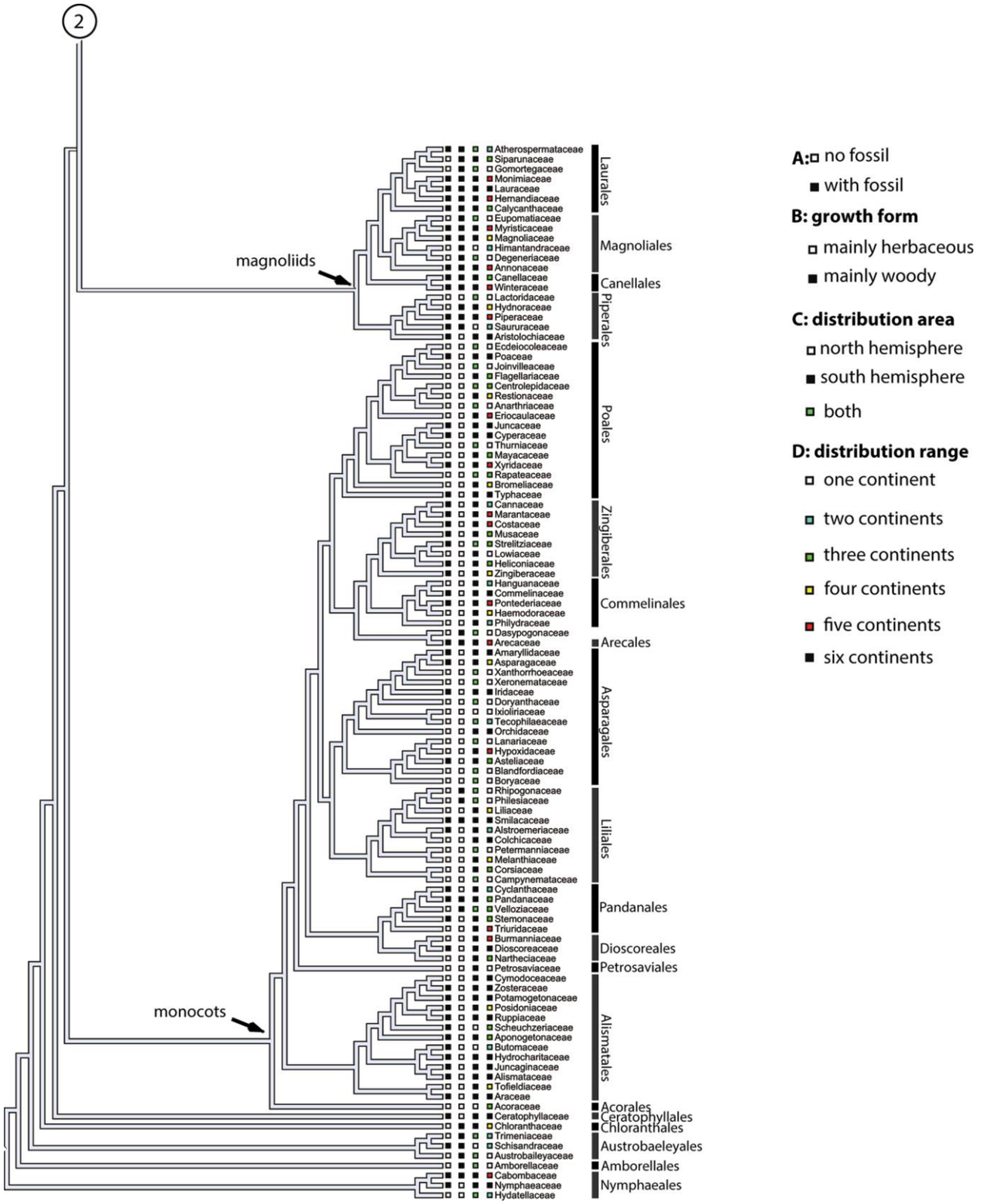


Fig. 4 (*Continued*)

Sampling Bias in the CAD

The numbers of fossil sites and assemblages vary substantially among the continents and between the epochs (table 1; fig. 3*B*–3*F*). Most Paleocene sites are from the central parts of North America and India. The majority of Eocene sites are from Europe, Asia, South America, and Australia. The number of North American sites in Oligocene beds is much lower than in Eocene beds, whereas an increase is evident in Europe and Asia. Globally, there are fewer Oligocene than Paleocene or Eocene sites. The European and Indian Miocene is very rich in fossil sites, but very few Miocene Antarctic sites are known. There are few Pliocene sites on any continent.

The number of assemblages for each epoch varies from 195 (Pliocene) to 946 (Miocene; table 1; fig. 3*B*–3*F*). There is no significant relationship between the number of sites and the number of assemblages and duration of the epoch (table 2; site number vs. duration: $R^2 = 0.26$, $P = 0.38$; assemblage number vs. duration: $R^2 = 0.30$, $P = 0.34$). Inspection of the data shows that the Paleocene and the Miocene are better sampled than the other three epochs.

The numbers of fossil sites and assemblages vary substantially among the continents (table 1). The Northern Hemisphere continents (i.e., Asia, Europe, and North America) have many more fossil sites and assemblages than the Southern Hemisphere continents (i.e., Africa, Antarctica, Oceania, and South America; site difference: $t = 3.56$, $P = 0.007$; assemblage difference: $t = 3.43$, $P = 0.009$). There is no significant relationship between the number of fossil sites and assemblages and the areas of continents and exposed sedimentary rocks ($R^2 = 0.25$–0.3; table 2).

The probability of a family being present in the fossil record is strongly predicted by the modern distribution of the family ($R^2 = 0.91$, $P < 0.01$; $\chi^2 = 51.0$, df = 5, $P < 0.01$), such that families now restricted to the Northern Hemisphere are significantly more likely to be in the fossil record than families currently restricted to the Southern Hemisphere (Fisher exact test, $P < 0.01$).

Of the 60 accepted angiosperm orders, 51 have at least one macrofossil record in the Cenozoic. The exceptions are Amborellales, Chloranthales, Acorales, Gunnerales, Picramniales, Berberidopsidales, Escalloniales, and Bruniales, which are all small and include only one to three extant families. At family level, only 221 of the current 445 families are represented in the fossil record (fig. 4; table 3). More than half of the families of the magnoliids, commelinids, and rosids are represented in the fossil record, whereas fewer than half of the families of the basal angiosperms, asterids, monocots (excluding commelinids), and eudicots (excluding rosids and asterids) are present in the fossil record (fig. 4).

Fossil record abundance varies between different growth forms and habitats. The mainly woody families, having nearly 30,000 occurrences, are more abundant than the herbaceous families, which have nearly 12,000 occurrences. Among the herbaceous families, families occurring in the wet habitat have more fossil records (ca. 9000 occurrences) than families in relatively dry habitats (ca. 2500 occurrences; fig. 5). Eight woody families have more than 1000 fossil records in the CAD, namely, the Betulaceae, Rosaceae, Lauraceae, Lythraceae, Fagaceae, Fabaceae, Juglandaceae, and Salicaceae (table 3). Among the herbaceous families, the Cyperaceae have the most abundant fossil record, usually preserved as fossil seeds. Aquatic herbaceous families such as the Typhaceae, Potamogetonaceae, and Nymphaeaceae also have relatively good fossil record (table 3).

Tests for phylogenetic signal indicate that the fossil record is phylogenetically constrained across the family-level phylogeny. Using the presence/absence data, the observed number of steps (119 steps) is significantly lower than the 95% bounds for step number after randomizing the terminal taxa (128–157 steps). Using the fossil abundance data (table 3), Blomberg's K is significantly larger than 0 (0.906; $P < 0.001$).

There is a significant evolutionary correlation between presence in the fossil record and continental distribution range, distribution on Northern/Southern Hemisphere, and woody/herbaceous habit (table 4). The maximum likelihood parameter estimates suggest that the evolutionary transition from "no fossil record" to "with fossil record" is 0.99 if lineages are inferred to be woody, versus 0.01 when they are herbaceous; 0.7 if they occur in the Northern Hemisphere, versus 0.41 in the Southern Hemisphere; and 0.46 if they occur on one continent, versus 0.51 if they occur on more than one continent. However, the latter are more striking when comparing occurrence on two or fewer continents versus more than two continents (transition to fossilization 0.45 on two continents vs. 1.5 when on more than two continents) or occurrence on three or fewer continents versus more than three continents (transition to fossilization 0.29 on three or fewer continents vs. 1.58 on more than three continents).

Completeness Test

The mean family-level SCM for the seven time intervals (early Eocene, middle Eocene, late Eocene, Oligocene, early Miocene, middle Miocene, late Miocene–Pleistocene) varies between 65% and 90% (fig. 6*A*), with the highest values for the Oligocene and the lowest for the middle and late Eocene (fig. 6*A*). Linear regression indicates that the SCM is significantly positively correlated with the number of assemblages for each time interval ($R^2 = 0.94$, $P < 0.01$).

The mean SCM for the seven large angiosperm groups is between 30% and 50% (fig. 6*B*). The rosids have the highest SCM values at family level, followed by the asterids, the commelinids, the magnoliids, and the eudicots (excluding rosids and asterids). The basal angiosperm and the monocots (excluding commelinids) have the lowest value for the SCM at 30% (fig. 6*B*).

The mean SCM for angiosperms from different habitats varies between 60% and 75%. The wetland herbaceous have the highest completeness (0.73), followed by the woody angiosperms (0.7) and nonaquatic herbaceous (0.62; fig. 5).

Discussion

As previously mentioned, the Cenozoic angiosperm fossil record is rich, and our results indicate that it could be used to obtain estimates of changes in angiosperm diversity through time, at least at generic and family levels. Nevertheless, our results show significant biases as well in the CAD, possibly also in Cenozoic paleobotany as a whole. The most obvious one is the interaction of time and continent: during some epochs, certain

Table 3

Occurrence Numbers of Each Angiosperm Group in the Cenozoic Angiosperm Database

Taxa	Occurrence no.
Basal angiosperm	**865**
Austrobaileyales	**26**
Illiciaceae	1
Schisandraceae	25
Nymphaeales	**839**
Cabombaceae	283
Nymphaeaceae	556
Magnoliids	**2481**
Canellales	**9**
Canellaceae	1
Winteraceae	8
Laurales	**1386**
Atherospermataceae	13
Calycanthaceae	2
Hernandiaceae	6
Lauraceae	1358
Monimiaceae	7
Magnoliales	**762**
Annonaceae	133
Magnoliaceae	620
Myristicaceae	9
Piperales	**324**
Aristolochiaceae	5
Piperaceae	22
Saururaceae	297
Commelinids	**4417**
Arecales	**399**
Arecaceae	399
Commelinales	**27**
Commelinaceae	7
Pontederiaceae	20
Poales	**3674**
Cyperaceae	2338
Juncaceae	30
Poaceae	304
Sparganiaceae	2
Typhaceae	994
Xyridaceae	6
Zingiberales	**317**
Cannaceae	4
Costaceae	1
Heliconiaceae	9
Marantaceae	9
Musaceae	8
Strelitziaceae	1
Zingiberaceae	294
Monocot (excluding commelinids)	**2916**
Alismatales	**2810**
Alismataceae	813
Aponogetonaceae	1
Araceae	766
Butomaceae	41
Hydrocharitaceae	338
Juncaginaceae	4
Posidoniaceae	1
Potamogetonaceae	829
Ruppiaceae	11
Scheuchzeriaceae	5
Zosteraceae	1

Table 3 (*Continued*)

Taxa	Occurrence no.
Asparagales	**17**
Amaryllidaceae	5
Asparagaceae	8
Hemerocallidaceae	1
Iridaceae	1
Laxmanniaceae	2
Orchidaceae	2
Dioscoreales	**11**
Dioscoreaceae	11
Liliales	**61**
Alstroemeriaceae	1
Asteliaceae	1
Ripogonaceae	3
Smilacaceae	56
Pandanales	**8**
Cyclanthaceae	5
Pandanaceae	5
Stemonaceae	3
Ceratophyllales	**158**
Ceratophyllales	**158**
Ceratophyllaceae	158
Eudicot (excluding rosids and asterids)	**3325**
Sabiaceae	130
Buxales	**26**
Buxaceae	26
Caryophyllales	**601**
Aizoaceae	1
Amaranthaceae	96
Caryophyllaceae	43
Droseraceae	201
Montiaceae	3
Nyctaginaceae	8
Phytolaccaceae	1
Polygonaceae	246
Sarcobataceae	1
Tamaricaceae	1
Dilleniales	**27**
Dilleniaceae	27
Proteales	**631**
Nelumbonaceae	43
Platanaceae	370
Proteaceae	218
Ranunculales	**773**
Berberidaceae	107
Eupteleaceae	3
Lardizabalaceae	13
Menispermaceae	223
Papaveraceae	69
Ranunculaceae	358
Santalales	**86**
Loranthaceae	12
Olacaceae	24
Santalaceae	43
Schoepfiaceae	7
Saxifragales	**963**
Cercidiphyllaceae	246
Crassulaceae	1
Grossulariaceae	52
Haloragaceae	189
Hamamelidaceae	230
Iteaceae	11
Saxifragaceae	4

Table 3 (*Continued*)

Taxa	Occurrence no.
Trochodendrales	**76**
Trochodendraceae	76
Asterids	**6260**
Apiales	**603**
Apiaceae	65
Araliaceae	525
Pittosporaceae	13
Aquifoliales	**149**
Aquifoliaceae	149
Asterales	**283**
Asteraceae	99
Campanulaceae	15
Menyanthaceae	169
Cornales	**939**
Cornaceae	852
Hydrangeaceae	62
Nyssaceae	25
Dipsacales	**797**
Adoxaceae	448
Caprifoliaceae	349
Ericales	**1882**
Actinidiaceae	189
Clethraceae	9
Cyrillaceae	2
Ebenaceae	95
Epacridaceae	8
Ericaceae	404
Lecythidaceae	28
Pentaphylacaceae	150
Polemoniaceae	2
Primulaceae	215
Sapotaceae	78
Styracaceae	211
Symplocaceae	385
Theaceae	106
Garryales	**169**
Eucommiaceae	75
Garryaceae	3
Icacinaceae	91
Gentianales	**374**
Apocynaceae	134
Gentianaceae	3
Loganiaceae	18
Rubiaceae	219
Lamiales	**977**
Acanthaceae	15
Bignoniaceae	47
Calceolariaceae	1
Lamiaceae	543
Oleaceae	208
Orobanchaceae	1
Paulowniaceae	6
Pedaliaceae	9
Plantaginaceae	122
Scrophulariaceae	10
Verbenaceae	15
Paracryphiales	**6**
Paracryphiaceae	6
Solanales	**81**
Convolvulaceae	6
Solanaceae	75

Table 3 (*Continued*)

Taxa	Occurrence no.
Rosids	**18,946**
Brassicales	**191**
Akaniaceae	2
Brassicaceae	32
Capparaceae	120
Cleomaceae	37
Celastrales	**103**
Celastraceae	103
Crossosomatales	**109**
Stachyuraceae	3
Staphyleaceae	104
Strasburgeriaceae	2
Cucurbitales	**124**
Coriariaceae	11
Corynocarpaceae	1
Cucurbitaceae	74
Datiscaceae	37
Tetramelaceae	1
Fabales	**1239**
Fabaceae	1233
Polygalaceae	6
Fagales	**5271**
Betulaceae	2055
Casuarinaceae	24
Fagaceae	1298
Juglandaceae	1045
Myricaceae	741
Nothofagaceae	106
Ticodendraceae	1
Geraniales	**1**
Geraniaceae	1
Huerteales	**3**
Dipentodontaceae	1
Tapisciaceae	2
Malpighiales	**2358**
Achariaceae	11
Calophyllaceae	26
Chrysobalanaceae	13
Clusiaceae	39
Dichapetalaceae	11
Elatinaceae	21
Erythroxylaceae	5
Euphorbiaceae	310
Humiriaceae	2
Hypericaceae	431
Irvingiaceae	1
Ixonanthaceae	1
Linaceae	5
Malpighiaceae	31
Ochnaceae	6
Passifloraceae	7
Peraceae	1
Phyllanthaceae	163
Putranjivaceae	4
Rhizophoraceae	15
Salicaceae	1021
Trigoniaceae	9
Violaceae	224
Malvales	**515**
Bixaceae	4
Cistaceae	2
Dipterocarpaceae	124
Malvaceae	349
Thymelaeaceae	36

Table 3 (*Continued*)

Taxa	Occurrence no.
Myrtales	**1954**
Combretaceae	147
Lythraceae	1342
Melastomataceae	49
Myrtaceae	207
Onagraceae	203
Vochysiaceae	6
Oxalidales	**203**
Connaraceae	16
Cunoniaceae	55
Elaeocarpaceae	44
Oxalidaceae	88
Rosales	**3971**
Cannabaceae	341
Elaeagnaceae	8
Moraceae	692
Rhamnaceae	427
Rosaceae	1724
Ulmaceae	536
Urticaceae	243
Sapindales	**2023**
Anacardiaceae	353
Burseraceae	73
Meliaceae	138
Rutaceae	300
Sapindaceae	986
Simaroubaceae	173
Vitales	**874**
Vitaceae	874
Zygophyllales	**7**
Zygophyllaceae	7
Unplaced	**80**
Boraginaceae	80

Note. Bold text indicates angiosperm orders or above-taxonomic levels.

continents have a rich record, while others are well represented during other epochs. Africa is generally poorly represented. There is also a phylogenetic bias. More than half of the families of the magnoliids, commelinids, and rosids are represented in the fossil record, whereas fewer than half of the families of the other basal angiosperms, the monocots (excluding commelinids) and eudicots (excluding rosids), are represented, even though collectively these latter clades are more species rich than the former group of clades. Furthermore, the angiosperm macrofossil record is biased by the growth form, habitat, and distribution of the families. Widespread families and families occurring in the Northern Hemisphere and in wet habitat are more likely to be collected and identified. Rosids have the most abundant fossil record, followed by the magnoliids and the basal angiosperms, in order of the number of occurrences. Incorrect identification is a persistent problem, but we currently have no indication how extensive it is. These biases should be taken into account when using the CAD.

Comparing with the Paleobiology Database (PaleoDB)

Currently, the PaleoDB is the most popular paleontological database, and it has a very complicated structure that allows it to include more comprehensive information. Compared with the PaleoDB, the CAD has a much simpler structure and is a more easily used interface. Besides this, it differs from the PaleoDB in several features: (1) the CAD is an assemblage-based database, while the PaleoDB is a collection-based database (in the PaleoDB, one assemblage could have several collections made by different groups); (2) the CAD includes several features not represented in the PaleoDB, such as the nearest living relatives, growth form of the fossil taxa, and so on; (3) the CAD uses the most recent classification for angiosperms, the APG III (Angiosperm Phylogeny Group 2009), while the PaleoDB uses the Cronquist system (Cronquist 1968); (4) the CAD has much better coverage for Africa, Asia, and Australia, while most of angiosperm fossil data in the PaleoDB come from North America and Europe.

Timescale Biases

Most approaches estimate the variation of species richness through time by counting taxa sampled in different time intervals (see reviews in Alroy 2010). For instance, Niklas and colleagues (Niklas et al. 1983; Niklas 1988; Niklas and Tiffney 1994) reconstructed the past diversity of the plants since the Upper Silurian using 29 geologic time intervals with significantly

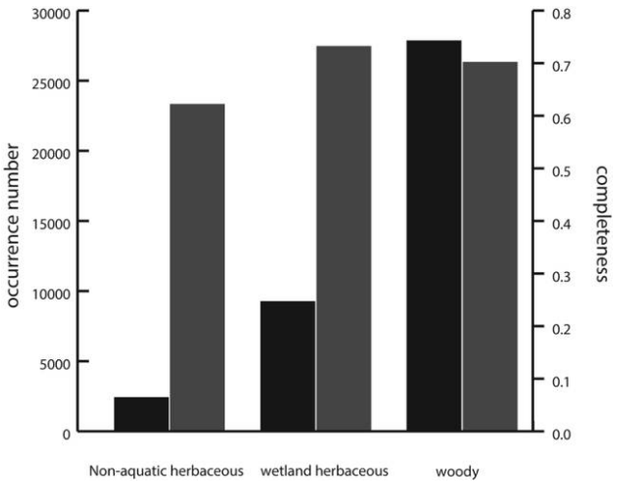


Fig. 5 The abundance and completeness of angiosperms of different growth forms and habitats. Black bars = occurrence number; gray bars = completeness.

different lengths. Although we could not demonstrate a significant relationship between the number of fossil sites and the duration of time intervals, positive relationships between richness and the duration of time intervals have been reported in many studies (e.g., Lloyd and Friedman 2013; Xing et al. 2014). Therefore, we recommend that equal-length time intervals should be used in diversity estimation or some corrections used to minimize any such sampling bias.

Continental biases. Sampling density varies among the continents. This bias has been documented for other organisms such as the Cretaceous tetrapods (Benson et al. 2013). This sampling bias is not accounted for by the continental size and exposed sedimentary rocks (table 2). Three factors may contribute to this bias.

1. *Collecting effort*. Paleobotanical studies have a much longer history in Europe and North America than in the Southern Hemisphere, and there are many more active paleobotanists in these northern countries than in the southern countries. The paleobotanical investigation of South America, Antarctica, and Oceania started only after the arrival of early explorers such as Darwin (Darwin and Bettany 1890) and Engelhardt (Engelhardt and Ochsenius 1891), who gathered fossils in early time (Nordenskjöld et al. 1905). In Australia, fossil sites are mostly along the southeast coast and in Tasmania and are almost absent from the central and northern parts. These areas are vast (two-thirds of Australia) and have very low population densities; consequently, low sampling effort may contribute to the paucity of documented fossil deposits. The lack of local paleobotanists may also account for the few paleobotanical studies in Southeast Asia, Mongolia, and Africa. Hopefully, current exploration will fill some of these gaps (Jacobs et al. 2010).

2. *Geological history*. Although Europe has abundant fossil sites, Paleocene and Eocene sites are much less common than in North America and Asia. This may be because large areas of the European platform were covered by epicontinental seas during the Paleogene, with continentalization increasing from the early Oligocene (Rögl 1999). By contrast, Neogene fossil sites are much less common in North America compared to Eurasia. This may be due to an expansion of grassland and sclerophyll vegetation, which are less likely to be preserved as macrofossils (Axelrod 1975; Retallack 1997). In the Southern Hemisphere, central and southern Africa were strongly influenced by mantle plumes (Storey 1995), which resulted in extensive uplift and doming and the formation of narrow coastal plains (Chorowicz 2005; De Wit 2007). Consequently, there is limited Cenozoic onshore sedimentation, contrasting with the massive Jurassic and Permian Karoo deposits. East African doming was followed by rifting and other igneous processes creating depositional basins, but because of seasonally dry climates and oxidation of sediments, these were mostly good for bone preservation and poor for organic matter. Geologically, the central and northern regions of Australia are less favorable for fossils because the landscape is very old, flat, and without volcanism and any significant tectonic activity during the Cenozoic (Veevers and Conaghan 1984; Raymond et al. 2007). As a result, there were few lakes formed by lava flows, landslips, faulting, and volcanic eruptions. Moreover, it is possible that relatively large areas of central and northern Australia were too dry for fossilization during much of the Cenozoic (Carpenter et al. 2014). The poor fossil record of Antarctica is due mainly to the presence of a permanent ice sheet since the Miocene (Zachos et al. 2001).

3. *Current geography*. Much of the surface area of South America is covered by dense forests (Amazonia) or is in terrain that is difficult to access (Andes), leaving relatively few areas of accessible exposed sediment. Most potential areas are in Patagonia, where desert landforms and extensive sedimentary rocks crop out. Dense forests may also limit fossil collection in southeast Asia and the African Congo basin. The desert re-

Table 4

BayesTraits Results to Test for Evolutionary Correlation between Occurrence in the Fossil Record and Growth Form (Woody/Herbaceous), Distribution (Northern/Southern Hemisphere), and Distribution Range (on One Continent/More Than One Continent)

Correlation	Independent	Dependent	Log BF	Conclusion
Fossil–growth form	−544.28	−535.47	17.61	Dependent
Fossil–Northern/Southern Hemisphere	−379.27	−371.89	14.76	Dependent
Fossil–one continent/more than one continent	−534.43	−502.85	63.17	Dependent

Note. The harmonic mean averaged over five independent runs is indicated. BF = Bayes factor. Log BF of 2–5 is interpreted as positive, 5–10 as strong, and >10 as highly significant support for the dependent model.

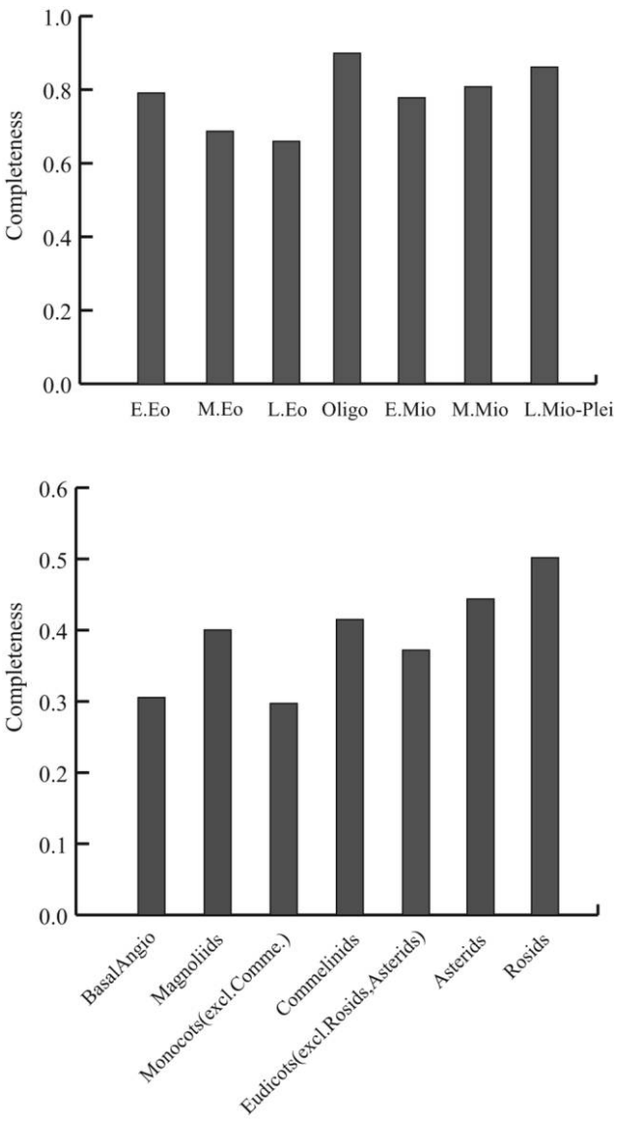


Fig. 6 The completeness of angiosperm fossil record. *Top*, completeness of angiosperm fossil record of seven time intervals in the Cenozoic. E. = early, M. = middle, L. = late, Eo = Eocene, Mio = Miocene, Plei = Pleistocene. *Bottom*, mean completeness of seven major groups of angiosperms.

gions such as central Australia, the African Sahara, and the Asian Gobi region have been subject to very deep weathering that may have destroyed many organic fossil deposits.

Phylogenetic Biases and Completeness

The phylogenetic bias of Cenozoic angiosperm fossil records is dependent on the taxonomic scale. There is no evident phylogenetic bias at the ordinal level, with most orders being represented in the fossil record. However, there is a strong phylogenetic bias at the family level. The families of the rosids have a better fossil record than the remainder of the angiosperms (fig. 4; table 3).

The causes of the phylogenetic biases may be complex (Behrensmeyer 1984; Kidwell and Flessa 1995; Kidwell and Holland 2002). Clearly, different types of organisms have different preservation potential (Tegelaar et al. 1989; Briggs 2003). In angiosperms, this applies primarily to different growth forms. Compared to woody species, herbaceous species have less lignin and consequently have a lower preservation potential. Also, the leaves of trees are more abundant, and transport from the plant to a site of deposition is more likely, whereas herbs tend to be preserved in autochthonous settings. Our data demonstrate that woody families (such as are very common in the rosids) are much better represented than herbaceous taxa (which are more common in the monocots and asterids; table 3; fig. 4). It is evident that herbaceous families in wet habitats are better represented than families in dry habitats. Also, widespread families have a better chance to be preserved in the fossil record, as do families occurring in the Northern Hemisphere. The phylogenetic correlation between occurrence in the fossil record and hemisphere, distribution range, and woodiness shows that these variables account for the phylogenetic signal in fossilization. Intrinsic as well as extrinsic characteristics of angiosperm families have affected preservation probability, and these characteristics are phylogenetically conserved. However, it is worth noting that clades with poor macrofossil record, such as the asterids and monocots, might have abundant microfossil record (Muller 1981).

Benton and Storrs (1994) regarded a completeness value of <10% as poor and >90% as excellent. The completeness for most of the time intervals during the Cenozoic is relatively good if only families with fossil records are considered. It is evident that the completeness of each time interval is positively correlated with the sampling density. The poorly sampled middle to late Eocene needs more attention in future paleobotanical studies. The completeness of angiosperm fossil record is also biased toward dry habitats. Almost no macrofossil record has been found for desert plants. As the Cenozoic angiosperm fossil record is phylogenetically biased, the mean completeness for seven main groups is rather low when families without fossil record are included. The completeness bias for different groups should be considered in future paleodiversity estimation.

Identification Error and Evaluation of Fossils

The study of the fossil record also needs to allow for potential identification error (Crepet 2008; Gandolfo et al. 2008). Determination of fossil species based on fragmentary organs is difficult. Before the 1980s, most fossil remains from Cenozoic sediments were described based on the premise that there is a unique most similar extant species for every given fossil (Wolfe 1973). This technique produced many fossil misidentifications, as well as a high proportion of incorrect taxonomic placements (Dilcher 1973; Gandolfo et al. 2008). However, we have no test to indicate the extent of identification error in the CAD. There are two indications of potential problems: (1) nearly 11% of all angiosperm occurrences appear only once in the CAD, which may reflect identification error, and (2) the CAD includes some families (e.g., Aizoaceae, Orobanchaceae, and Crassulaceae) that were thought not to have a confirmed macrofossil record.

Reevaluation of identifications of all records in the CAD is not a trivial task. Here, we propose two options to improve the accuracy of identifications or reduce identification bias. The first

option is the ongoing process of critical taxonomic investigation of particular groups by experts, especially of potential problem taxa. Simply ignoring potential problem taxa may miss important information. Such focused critical investigations have been applied to several important critical taxa such as the Asteridae (Martínez-Millán 2010), Casuarinaceae (Zamaloa et al. 2006), Ericaceae subfamily Styphelioideae (Jordan and Hill 1996), *Eucalyptus* (Hermsen et al. 2012), fin-winged fruit fossils (Manchester and O'Leary 2010), Malvaceae (Worobiec et al. 2010), Menispermaceae (Jacques et al. 2007), Proteaceae (Carpenter 2012), and Rubiaceae (Graham 2009). More critical investigations of potentially misidentified taxa are encouraged in the future. An alternative option is working at several taxonomic levels, especially in studies such as paleodiversity and diversification rate estimations (Raup 1972; Niklas and Tiffney 1994) to get cross validation.

In addition to identification errors, special rules for names of fossils might be another source of bias in paleodiversity estimation. Due to the fragmentation of fossil preservation, fossils of particular parts, life-history stages, or preservational states were usually given different names (morphotaxa) even though they belong to the same organism (McNeill and Turland 2011). This bias should also be considered in future diversity estimations. One realistic solution to this problem is to use only the fossil taxa from one plant part per particular plant group, as suggested by Cleal et al. (2012).

Conclusion

The Cenozoic macrofossil record of angiosperms is remarkably rich, especially that of woody families found in the Northern Hemisphere. As in any fossil record, there are numerous biases in these data. However, a global comparison of the fossil record reveals where the biases are, spatially, temporally, and phylogenetically. Even if incorrect identifications are rampant, this database would serve as a useful tool for researchers working on the history of particular groups because they could then easily track references, evaluate the fossils, and update the identifications. Consequently, a judicious use of the database should be highly informative.

Acknowledgments

We thank contributors of the Paleobiology Database for inputting data used in the CAD and T. Utescher and V. Mosbrugger for sharing data from Palaeoflora Database. We thank P. Wilf, V. Mosbrugger, and D. Uhl for their useful comments on the structure of the CAD. We also thank T. Li for developing the CAD website. The study is supported by the Swiss National Science Foundation (grant 31003A_130847 to H. P. Linder) and US National Science Foundation (grants DEB 0918932 and DEB 0830020 to M. A. Gandolfo).

Literature Cited

Adrain JM, SR Westrop 2003 Paleobiodiversity: we need new data. Paleobiology 29:22–25.

Alroy J 2010 Fair sampling of taxonomic richness and unbiased estimation of origination and extinction rates. Quant Methods Paleobiol Paleontol Soc Pap 16:55–80.

Amiotte Suchet P, JL Probst, W Ludwig 2003 Worldwide distribution of continental rock lithology: implications for the atmospheric/soil CO_2 uptake by continental weathering and alkalinity river transport to the oceans. Glob Biogeochem Cycles 17:1038.

Angiosperm Phylogeny Group 2009 An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc 161:105–121.

Axelrod DI 1975 Evolution and biogeography of Madrean-Tethyan sclerophyll vegetation. Ann Mo Bot Gard 62:280–334.

Behrensmeyer AK 1984 Taphonomy and the fossil record: the complex processes that preserve organic remains in rocks also leave their own traces, adding another dimension of information to fossil samples. Am Sci 72:558–566.

Behrensmeyer AK, SM Kidwell 1985 Taphonomy's contributions to paleobiology. Paleobiology 11:105–119.

Benson RBJ, PD Mannion, RJ Butler, P Upchurch, A Goswami, SE Evans 2013 Cretaceous tetrapod fossil record sampling and faunal turnover: implications for biogeography and the rise of modern clades. Palaeogeogr Palaeoclimatol Palaeoecol 372:88–107.

Benton MJ 1987 Mass extinctions among families of non-marine tetrapods: the data. Mem Soc Geol Fr 150:21–32.

Benton MJ, AM Dunhill, GT Lloyd, FG Marx 2011 Assessing the quality of the fossil record: insights from vertebrates. Geol Soc Lond Spec Publ 358:63–94.

Benton MJ, GW Storrs 1994 Testing the quality of the fossil record: paleontological knowledge is improving. Geology 22:111–114.

Benton MJ, M Wills, R Hitchin 2000 Quality of the fossil record through time. Nature 403:534–537.

Blatt H, RL Jones 1975 Proportions of exposed igneous, metamorphic, and sedimentary rocks. Geol Soc Am Bull 86:1085–1088.

Blomberg SP, T Garland Jr, AR Ives 2003 Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745.

Briggs DE 2003 The role of decay and mineralization in the preservation of soft-bodied fossils. Annu Rev Earth Planet Sci 31:275–301.

Burnham RJ, RA Spicer 1986 Forest litter preserved by volcanic activity at El Chichon, Mexico: a potentially accurate record of the pre-eruption vegetation. Palaios 1:158–161.

Carpenter RJ 2012 Proteaceae leaf fossils: phylogeny, diversity, ecology and austral distributions. Bot Rev 78:261–287.

Carpenter RJ, SP McLaughlin, RS Hill, KJ McNamara, GJ Jordan 2014 Early evidence of xeromorphy in angiosperms: stomatal encaryption in a new Eocene species of *Banksia* (Proteaceae) from Western Australia. Am J Bot 101:1486–1497.

Cascales-Miñana B 2011 New insights into the reading of Paleozoic plant fossil record discontinuities. Hist Biol 23:115–130.

Cascales-Miñana B, CJ Cleal 2013 What is the best way to measure extinction? reflection from the palaeobotanical record. Earth-Sci Rev 124:126–147.

Cascales-Miñana B, JB Diez 2012 The effect of singletons and interval length on interpreting diversity trends from the palaeobotanical record. Palaeontol Electron 15:6A.

Cascales-Miñana B, J Muñoz-Bertomeu, R Ros, J Segura 2010 Trends and patterns in the evolution of vascular plants: macroevolutionary implications of a multilevel taxonomic analysis. Lethaia 43:545–557.

Chorowicz J 2005 The East African rift system. J Afr Earth Sci 43:379–410.

Cleal CJ, D Uhl, B Cascales-Miñana, BA Thomas, AR Bashforth, SC King, EL Zodrow 2012 Plant biodiversity changes in Carboniferous tropical wetlands. Earth-Sci Rev 114:124–155.

Crane PR, S Lidgard 1989 Angiosperm diversification and paleolatitudinal gradients in Cretaceous floristic diversity. Science 246:675–678.

Cronquist A 1968 The evolution and classification of flowering plants. New York Botanical Garden, Bronx, NY.

Crepet WL 2008 The fossil record of angiosperms: requiem or renaissance? Ann Mo Bot Gard 95:3–33.

Darwin C, GT Bettany 1890 Journal of researches into the natural history and geology of the countries visited during the voyage of *HMS Beagle* round the world: under the command of Capt. Fitz Roy, RN. Ward, Lock.

De Wit M 2007 The Kalahari Epeirogeny and climate change: differentiating cause and effect from core to space. S Afr J Geol 110: 367–392.

Dilcher D 1973 Revision of the Eocene flora of southeastern North America. Palaeobotanist 20:7–18.

Durham JW 1967 The incompleteness of our knowledge of the fossil record. J Paleontol 41:559–565.

Engelhardt H, KC Ochsenius 1891 Ueber Tertiärpflanzen von Chile. Commission bei M. Diesterweg.

Fara E, MJ Benton 2000 The fossil record of Cretaceous tetrapods. Palaios 15:161–165.

Foote M, JS Crampton, AG Beu, CS Nelson 2015 Aragonite bias, and lack of bias, in the fossil record: lithological, environmental, and ecological controls. Paleobiology 41:245–265.

Foote M, JJ Sepkoski 1999 Absolute measures of the completeness of the fossil record. Nature 398:415–417.

Fröbisch J 2013 Vertebrate diversity across the end-Permian mass extinction: separating biological and geological signals. Palaeogeogr Palaeoclimatol Palaeoecol 372:50–61.

Gandolfo MA, KC Nixon, WL Crepet 2008 Selection of fossils for calibration of molecular dating models. Ann Mo Bot Gard 95:34–42.

Gradstein FM, G Ogg, M Schmitz 2012 The geologic time scale 2012. 2 vols. Elsevier, Oxford.

Graham A 2009 Fossil record of the Rubiaceae. Ann Mo Bot Gard 96:90–108.

Greenwood DR 1991 The taphonomy of plant macrofossils. Pages 141–169 *in* SK Donovan, ed. The processes of fossilization. Columbia University Press, New York.

Hermsen EJ, MA Gandolfo, M del Carmen Zamaloa 2012 The fossil record of *Eucalyptus* in Patagonia. Am J Bot 99:1356–1374.

Jacobs BF, AD Pan, CR Scotese 2010 A review of the Cenozoic vegetation history of Africa. Pages 57–72 *in* L Werdelin, ed. Cenozoic mammals of Africa. University of California Press, Berkeley.

Jacques FM, C Gallut, R Vignes-Lebbe, RZ Bagils 2007 Resolving phylogenetic reconstruction in Menispermaceae (Ranunculales) using fossils and a novel statistical test. Taxon 56:379–392.

Jordan G, R Hill 1996 The fossil record of the Epacridaceae. Ann Bot 77:341–346.

Karr JA, ME Clapham 2015 Taphonomic biases in the insect fossil record: shifts in articulation over geologic time. Paleobiology 41:16–32.

Kidwell SM, KW Flessa 1995 The quality of the fossil record: populations, species, and communities. Annu Rev Ecol Syst 26:269–299.

Kidwell SM, SM Holland 2002 The quality of the fossil record: implications for evolutionary analyses. Annu Rev Ecol Syst 33:561–588.

Lidgard S, PR Crane 1988 Quantitative analyses of the early angiosperm radiation. Nature 331:344–346.

——— 1990 Angiosperm diversification and Cretaceous floristic trends: a comparison of palynofloras and leaf macrofloras. Paleobiology 16: 77–93.

Lloyd GT, M Friedman 2013 A survey of palaeontological sampling biases in fishes based on the Phanerozoic record of Great Britain. Palaeogeogr Palaeoclimatol Palaeoecol 372:5–17.

Lupia R, S Lidgard, PR Crane 1999 Comparing palynological abundance and diversity: implications for biotic replacement during the Cretaceous angiosperm radiation. Paleobiology 25:305–340.

Manchester SR, EL O'Leary 2010 Phylogenetic distribution and identification of fin-winged fruits. Bot Rev 76:1–82.

Martínez-Millán M 2010 Fossil record and age of the Asteridae. Bot Rev 76:83–135.

Maxwell WD, MJ Benton 1990 Historical tests of the absolute completeness of the fossil record of tetrapods. Paleobiology 16:322–335.

McNeill J, NJ Turland 2011 Major changes to the code of nomenclature—Melbourne, July 2011. Taxon 60:1495–1497.

Muller J 1981 Fossil pollen records of extant angiosperms. Bot Rev 47: 1–142.

Niklas KJ 1988 Patterns of vascular plant diversification in the fossil record: proof and conjecture. Ann Mo Bot Gard 75:35–54.

Niklas KJ, BH Tiffney 1994 The quantification of plant biodiversity through time. Philos Trans R Soc B 345:35–44.

Niklas KJ, BH Tiffney, AH Knoll 1983 Patterns in vascular land plant diversification. Nature 303:614–616.

Nordenskjöld O, JG Andersson, CJ Skottsberg, CA Larsen 1905 Antarctica: or, two years amongst the ice of the South Pole. Hurst & Blackett, London.

Pagel M, A Meade 2006 Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am Nat 167:808–825.

Paton AJ, N Brummitt, R Govaerts, K Harman, S Hinchcliffe, B Allkin, EN Lughadha 2008 Towards Target 1 of the Global Strategy for Plant Conservation: a working list of all known plant species—progress and prospects. Taxon 57:602–611.

Pearson MR, RBJ Benson, P Upchurch, J Fröbisch, CF Kammerer 2013 Reconstructing the diversity of early terrestrial herbivorous tetrapods. Palaeogeogr Palaeoclimatol Palaeoecol 372:42–49.

Qian H, J Zhang 2014 Using an updated time-calibrated family-level phylogeny of seed plants to test for non-random patterns of life forms across the phylogeny. J Syst Evol 52:423–430.

Raup DM 1972 Taxonomic diversity during the Phanerozoic. Science 177:1065–1071.

Raup DM, JJ Sepkoski Jr 1982 Mass extinctions in the marine fossil record. Science 215:1501–1503.

Raymond O, S Liu, P Kilgour, A Retter, A Stewart, G Stewart 2007 Surface geology of Australia, 1 : 1,000,000 scale. Geoscience Australia, Canberra.

Retallack GJ 1997 Neogene expansion of the North American prairie. Palaios 12:380–390.

Revell LJ 2013 Phylogenetic tools for comparative biology (and other things). R-package: phytools. https://cran.r-project.org/web/packages/phytools/index.html.

Rögl F 1999 Mediterranean and Paratethys: facts and hypotheses of an Oligocene to Miocene paleogeography (short overview). Geol Carpath 50:339–349.

Romero EJ 1986 Paleogene phytogeography and climatology of South America. Ann Mo Bot Gard 73:449–461.

Ronov AB 1982 The earth's sedimentary shell (quantitative patterns of its structure, compositions, and evolution): the 20th VI Vernadskiy Lecture, March 12, 1978. Int Geol Rev 24:1313–1363.

Sepkoski JJ, RK Bambach, DM Raup, JW Valentine 1981 Phanerozoic marine diversity and the fossil record. Nature 293:435–437.

Spicer RA 1991 Plant taphonomic processes: taphonomy, releasing the data locked in the fossil record. Top Geobiol 9:71–113.

Storey BC 1995 The role of mantle plumes in continental breakup: case histories from Gondwanaland. Nature 377:301–308.

Tegelaar E, J De Leeuw, S Derenne, C Largeau 1989 A reappraisal of Kerogen Formation. Geochim Cosmochim Acta 53: 3103–3106.

Utescher T, V Mosbrugger 2012 Palaeoflora database. http://www2.geo.uni-bonn.de/palaeoflora/palaeoflora_home.htm.

Veevers JJ, PJ Conaghan 1984 Phanerozoic earth history of Australia. Oxford University Press, New York.

Wolfe JA 1973 Fossil forms of Amentiferae. Brittonia 25:334–355.

Worobiec G, E Worobiec, Z Kvaček 2010 Neogene leaf morphotaxa of Malvaceae s.l. in Europe. Int J Plant Sci 171:892–914.

Xing Y, RE Onstein, RJ Carter, T Stadler, HP Linder 2014 Fossils and large molecular phylogeny show that the evolution of species richness, generic diversity and turnover rates are disconnected. Evolution 68:2821–2832.

Zachos J, M Pagani, L Sloan, E Thomas, K Billups 2001 Trends, rhythms, and aberrations in global climate 65 Ma to present. Science 292:686–693.

Zamaloa MC, MA Gandolfo, CC González, EJ Romero, NR Cúneo, P Wilf 2006 Casuarinaceae from the Eocene of Patagonia, Argentina. Int J Plant Sci 167:1279–1289.